Automatic Construction of UMLS Metathesaurus with Deep Learning

Joey Yip
Olivier Bodenreider

U.S. National Library of Medicine

Unified Medical Language System (UMLS) Metathesaurus

- Started in 1986 by the National Library of Medicine (NLM)

Overcome barriers to effective retrieval of machine-readable information

- The variety of ways the same concepts are expressed by different terminologies

(MeSH, MedDRA, RxNORM, ICD-10, SNOMED CT, etc)

- **~ 10** million English medical terms
- From **210** source vocabularies
  - **General**
    - *Anatomy (FMA, Neuronames), drugs (RxNorm, ATC, First DataBank), medical devices (UMD, SPN), clinical terms (SNOMED CT), information sciences (MeSH), administrative terminologies (ICD-9/10)*
  - **Specialized**
    - *Nursing (NIC), psychiatry (DSM, APA), adverse reactions (MedDRA)*
- Grouped into **~ 3.85** million concepts

Used in areas such as patient care, clinical coding, information retrieval, knowledge exploration, and data mining

Integrating Subdomains

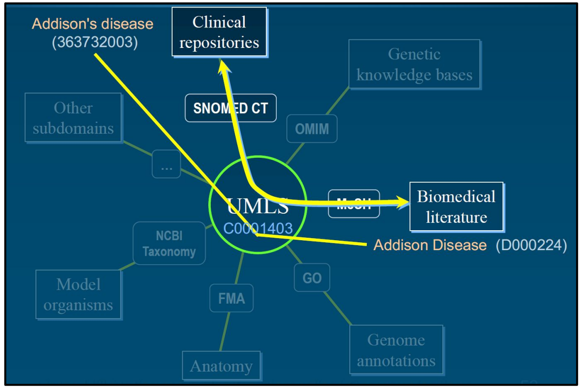


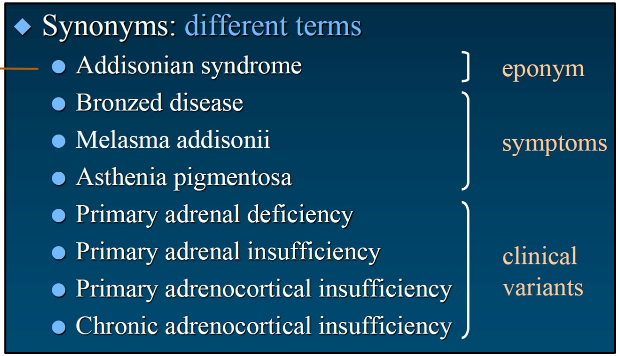
Image from ***Unified Medical Language System Overview*** *by Olivier Bodenreider*

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

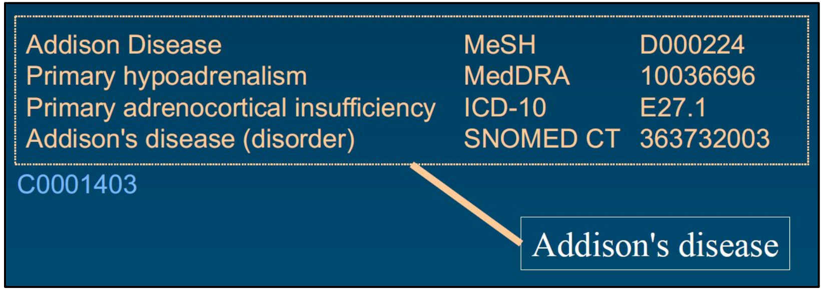
Unified Medical Language System (UMLS)

Addison’s Disease (Concept)

atom



Synonymous atoms are clustered into a concept with a **UMLS Concept Unique Identifier (CUI)**

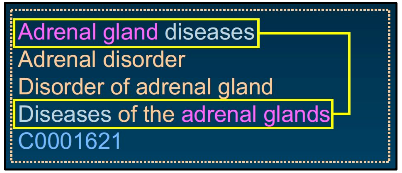


Images from ***Unified Medical Language System Overview***
by Olivier Bodenreider

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.

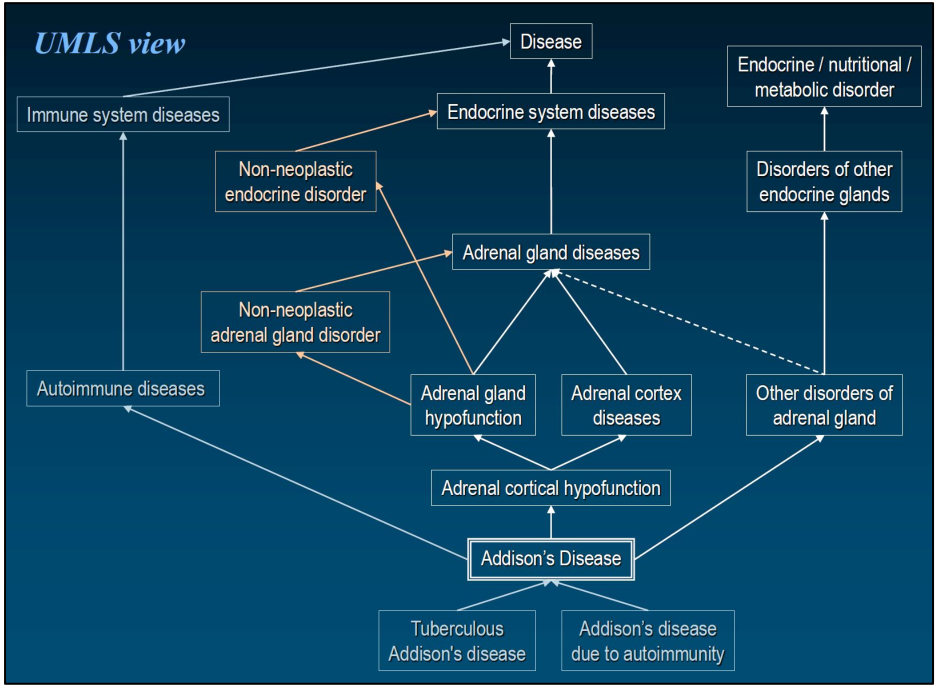
Construction of UMLS Metathesaurus *(Updates bi-annually)*

- **Lexical Knowledge**
  - *Lexical Variant Generator (LVG)*



adrenal disease gland

- **Semantic Pre-processing**
- **UMLS Human Editors**



Images from ***Unified Medical Language System Overview*** *by Olivier Bodenreider*

Motivation

The current approach in adding new resources from identifying lexical variants to manual audits can be both arduous and time-consuming.

*(~ **10** million English medical terms, ~ **3.85** million concepts)*

Objectives

The project explores the realm of *supervised machine learning approach (Deep Learning)* to

1. Identify **synonymy** and **non-synonymy** among UMLS concepts at the atom level
    - Given two atoms, are they synonymous (same CUI)?
2. Investigate Deep Learning approach could emulate the current building process

Problem Formulation

~~**Approach 1 (Classification task):**~~

- ~~**Training Data:** ~ 10M English language atoms and each with its own CUI assignment~~
- ~~We can train a classification model to predict which CUI should be assigned to a given "new" atom *(since atoms having the same CUI are synonymous).*~~
- ~~Input: Atom -> Output label: CUI~~
- ~~**Challenge:** ~ 3.85M softmax outputs (extreme classification task)~~

Approach 2 (Similarity task):

- Learn **similarities** between atoms within a CUI and **dissimilarities** between atoms from different CUIs.

A fully-trained model should identify and learn scenarios where

Lung disease and disorder	Two atoms that are ***lexically similar*** in nature but ***are not synonymous***	**Head** disease and disorder
Addison's disease	Two atoms that are ***lexically dissimilar*** but are ***synonymous***	Primary adrenal deficiency

Traditional Neural Network Architecture

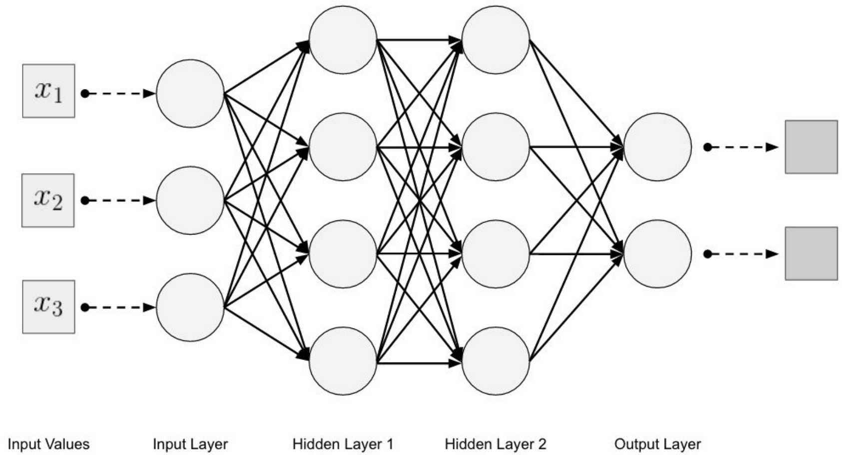


Image source: https://www.oreilly.com/library/view/deep-learning/9781491924570/ch04.html

Feedforward Neural Network (Multilayer Perceptron):

Not suited for Pairwise-similarity task

Siamese (Twin) Neural Network

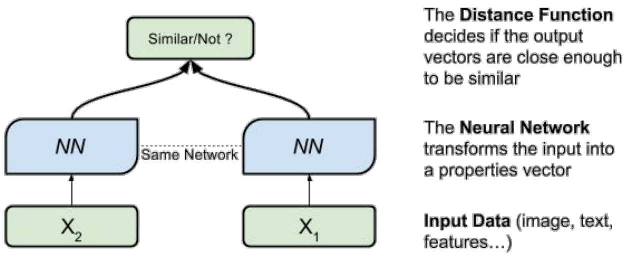
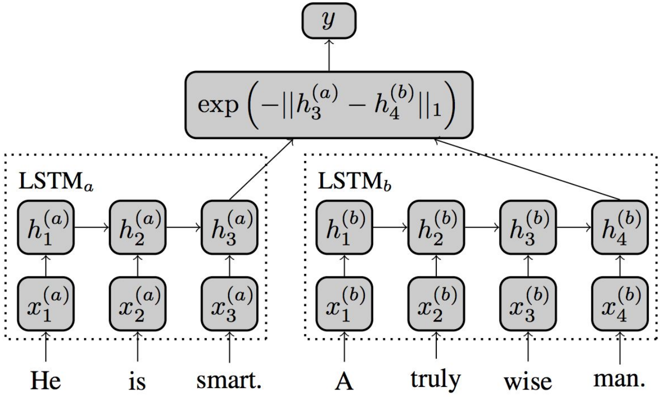


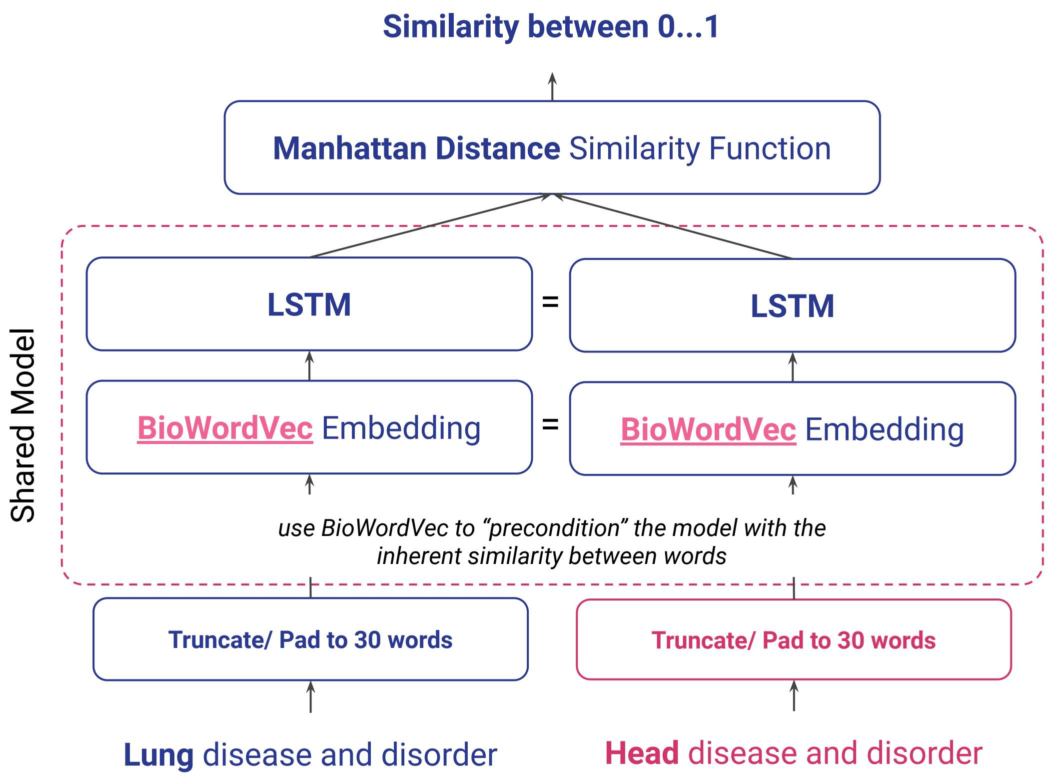
Image source:
https://aws.amazon.com/blogs/machine-learning/combining-deep-learning-networks-gan-and-siamese-to-generate-high-quality-life-like-images/

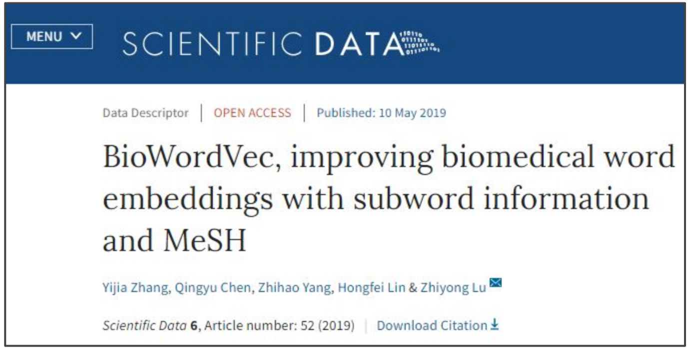


Mueller, J., & Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.

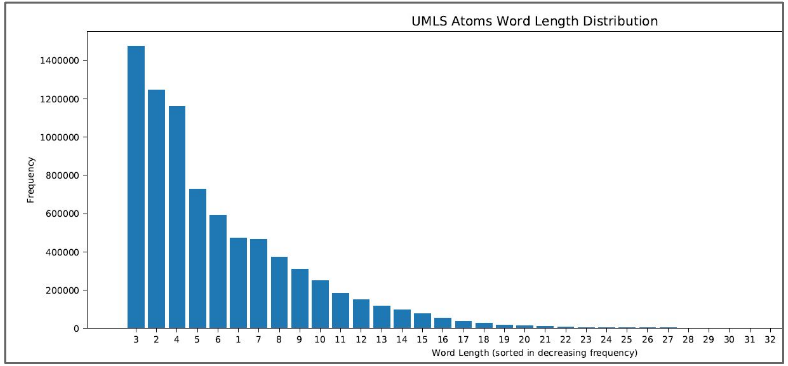
Image source: Siamese Recurrent Architectures for Learning Sentence Similarity

Siamese-LSTM




MENU
SCIENTIFIC DATA
Data Descriptor | OPEN ACCESS | Published: 10 May 2019
BioWordVec, improving biomedical word embeddings with subword information and MeSH
Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin & Zhiyong Lu
Scientific Data 6, Article number: 52 (2019) | Download Citation

Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific Data. 2019.



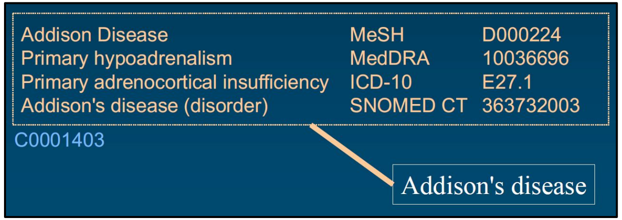
UMLS Atoms Word Length Distribution
(Word length 30 covers 97% of atoms in the UMLS)

Mueller, J., & Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.

Dataset (2019-AA UMLS) and Feature Engineering

Positive Pairs (Synonyms)

- (CUI)-asserted synonymy between atoms (~15 million pairs)



Negative Pairs (Non-synonyms)

- Ideally, we want to generate all negative pairs (1 atom against atoms from other non-related CUIs) (~ 10 million atoms * 10 million atoms)

Class Imbalance: Number of *Non-synonyms* > Number of *Synonyms*

Intuition: What we want are interesting negative pairs that are <u>lexically similar but differ in semantics</u>.

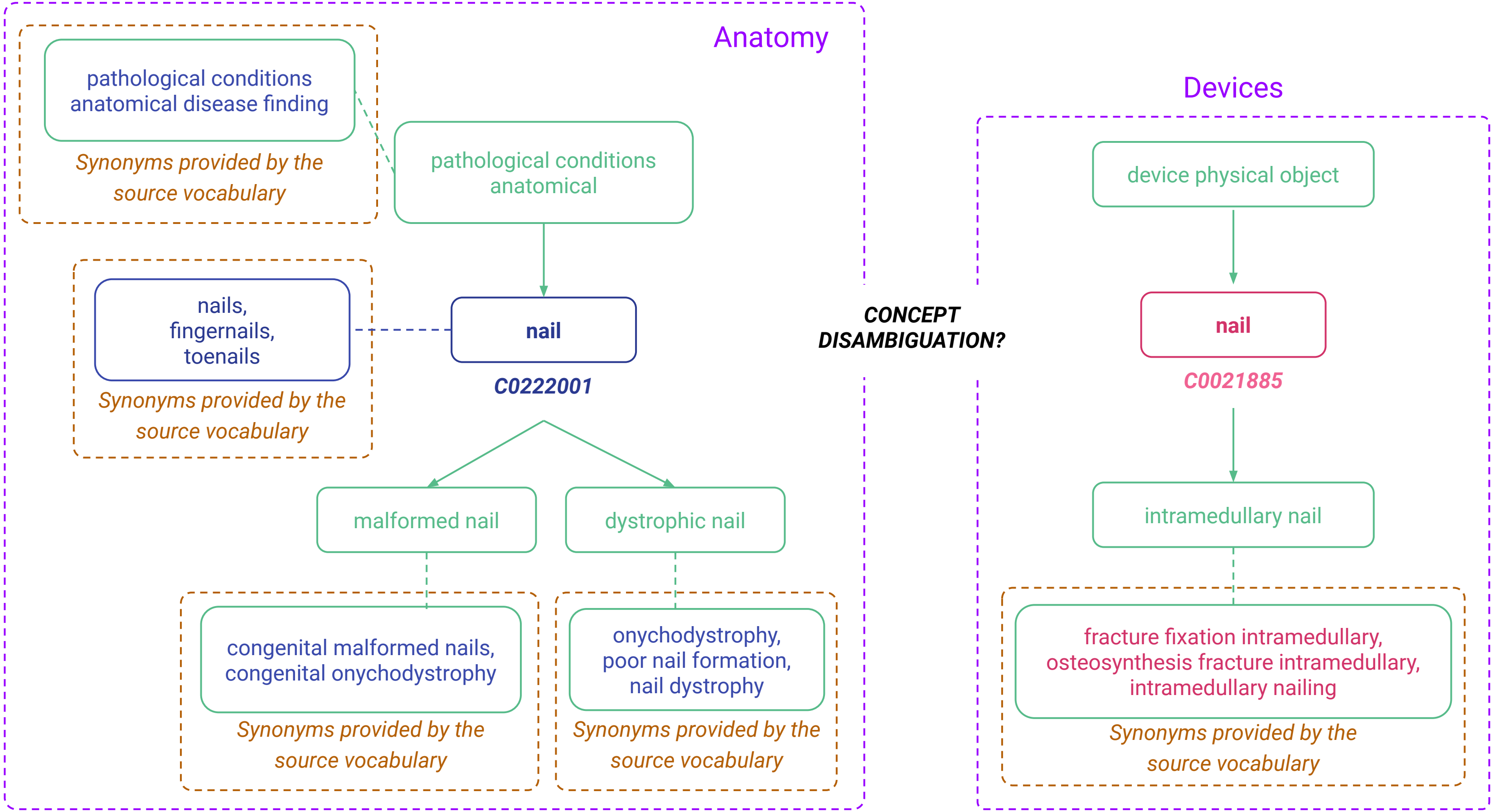
- Heuristic Approach: Use **Jaccard Index** to generate **negative pairs** for atoms with **high Jaccard Similarity** *(Sort and filter top ~15 million pairs)*

C0000473:
Product containing ***para-aminobenzoic*** acid

C0038784:
Product containing ***sulfuric*** acid

Intersection (3)/ Union (5)
= **0.6 Jaccard Index**

Going beyond atoms... Let's Contextualize!



Experimental Setup

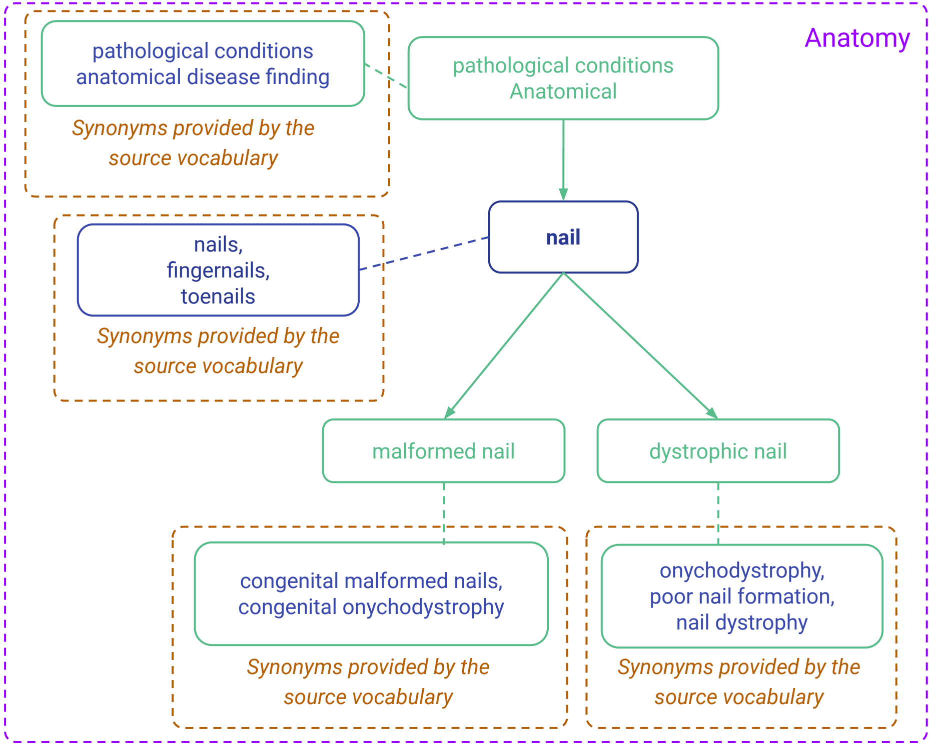
1. “Base” experiment
(Atom lexical features)

2. “Base” (Atom lexical features)
+ Synonyms provided by the source vocabulary

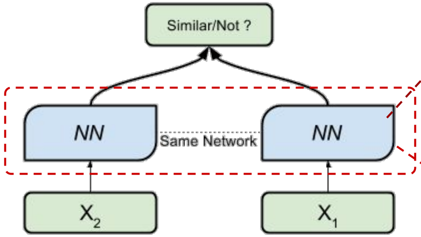
3. “Base” (Atom lexical features)
+ *Hierarchical-Context(atom)*
+ *Semantic Group*

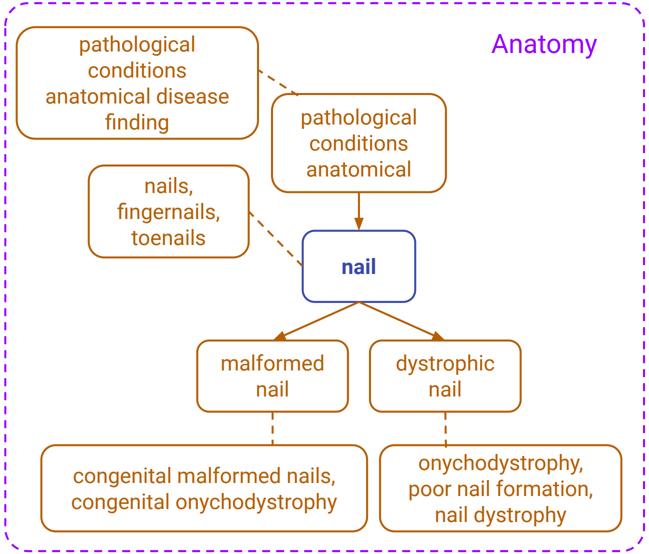
4. “Base” (Atom lexical features)
+ *Synonyms provided by the source vocabulary*
+ *Hierarchical-Context(atom)*
+ *Semantic Group*

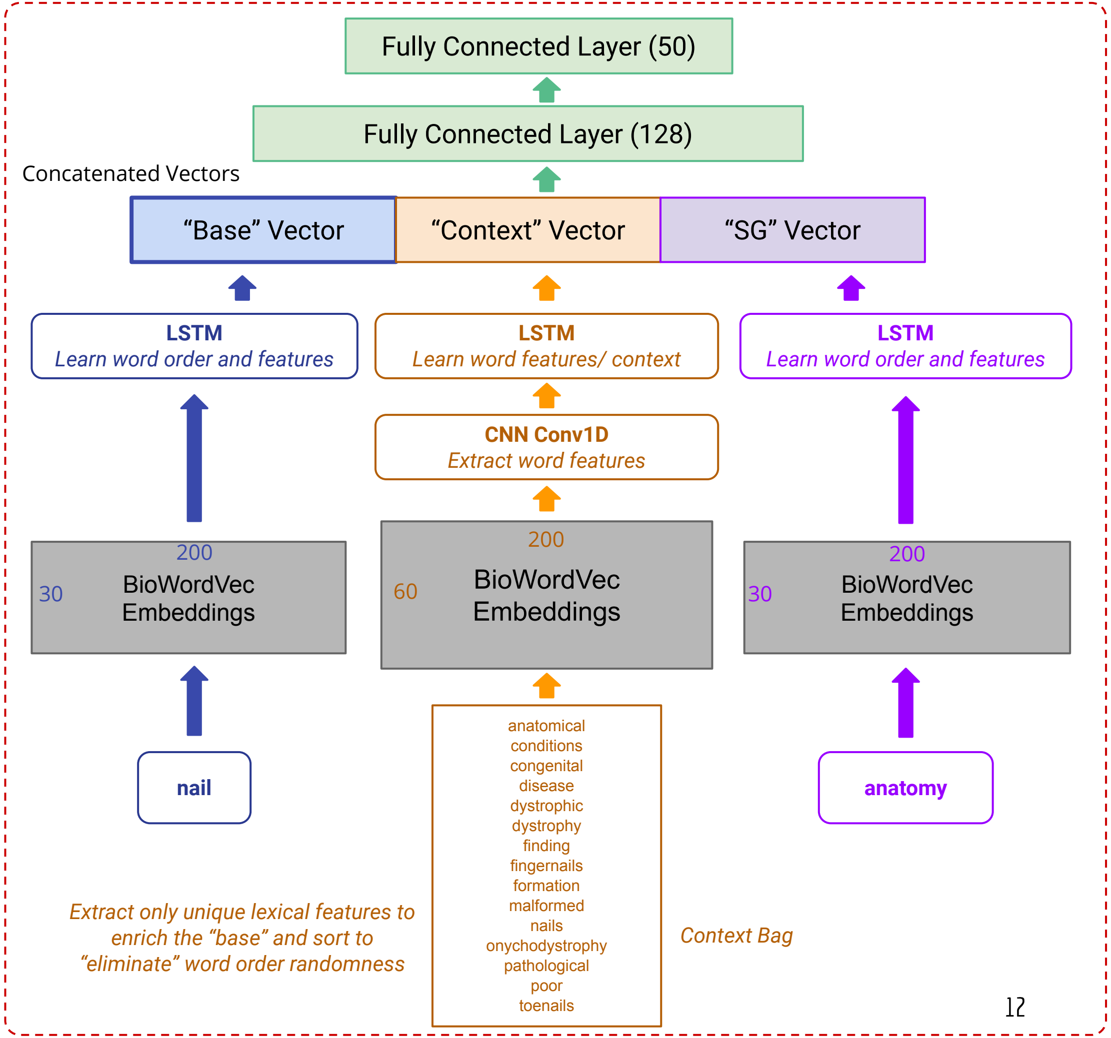
5. “Base” (Atom lexical features)
+ *Synonyms provided by the source vocabulary*
+ *Hierarchical-Context(atom)*
+ *Synonyms of the Hierarchical-Context(atom)*
+ *Semantic Group*



Architecture

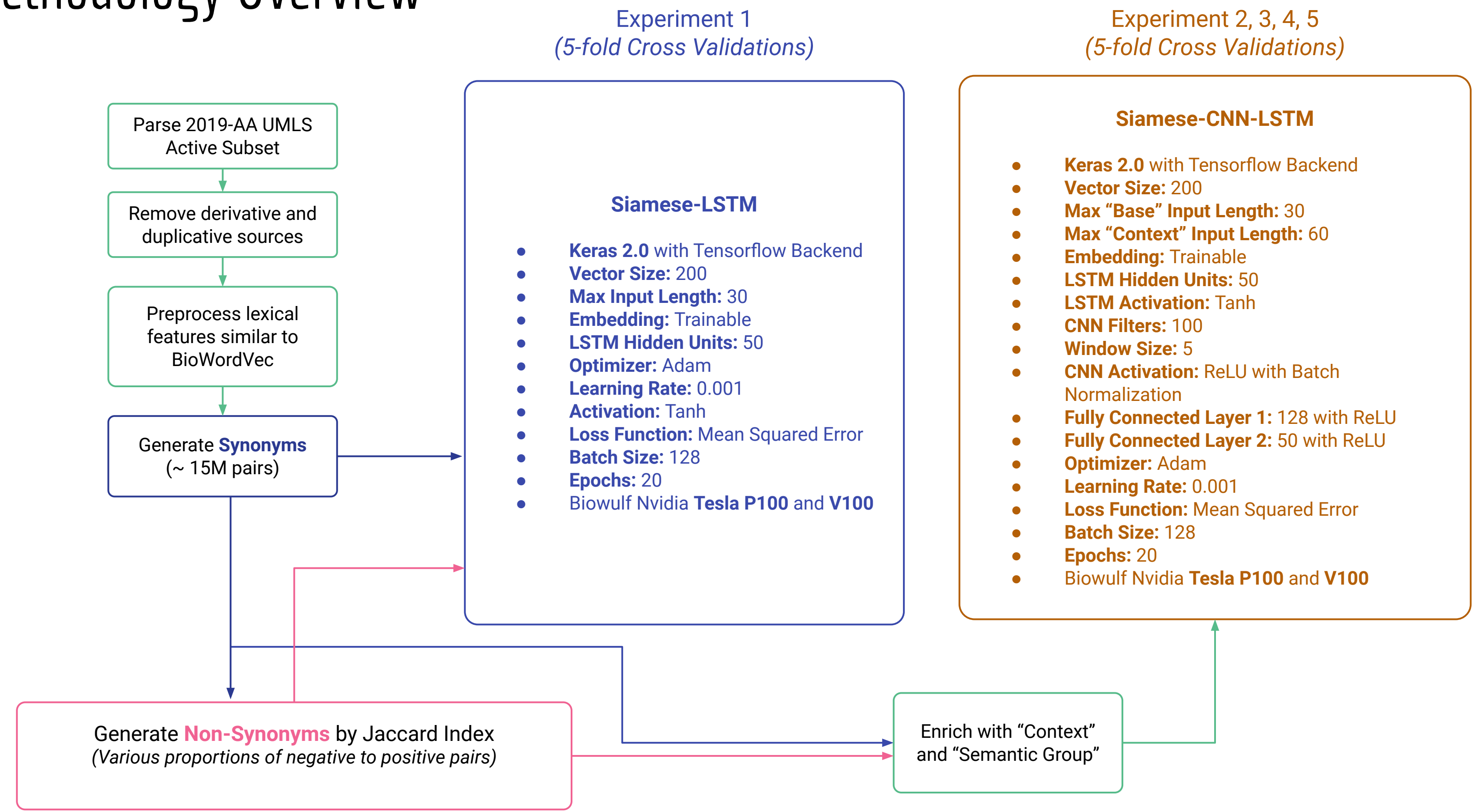






Pontes, E. L., Huet, S., Linhares, A. C., & Torres-Moreno, J. M. (2018). Predicting the Semantic Textual Similarity with Siamese CNN and LSTM. arXiv preprint arXiv:1810.10641.

Methodology Overview



Results & Evaluations *(based on optimal runs)*

Model/ Performance Metrics	Base	Base + Source Synonymy	Base + Hier. Context + Semantic Group	Base + Source Synonymy + Hier. Context + Semantic Group	Base + Source Synonymy + Hier. Context + Hier. Source Synonymy + Semantic Group
Accuracy	0.9333	0.8720	0.9486	0.9520	**0.9541**
Precision	0.7828	**0.8654**	0.7643	0.8296	0.8009
Recall	0.7379	0.8874	0.8381	**0.9038**	0.8978
F1-Score	0.7597	**0.8763**	0.7995	0.8428	0.8466
Matthew CC	0.7214	0.7441	0.7712	**0.8173**	0.8215
Specificity	0.9659	0.8560	**0.9640**	0.9601	0.9633
Sensitivity	0.7379	0.8874	0.8381	**0.9038**	0.8978
False Positive Rate	**0.0341**	0.1440	0.0360	0.0399	0.0367

Observations:

- Source synonymy is responsible for achieving high precision and overall F-1 score.
- Adding hierarchical context trades precision for higher recall.
- Adding source synonymy, hierarchical context, and semantic group give an overall boost to accuracy and recall.
- However, adding source synonymy of hierarchical context did not yield any noticeable improvement.

Examples of True Positives and True Negatives Correctly Identified

True Positives (Synonyms) Correctly Identified	
nail clipper	cutters nail
injury of salivary gland	salivary gland injury
avulsion	fracture sprain
True Negatives (Non-synonyms) Correctly Identified	
fingernail	infection of fingernail
product containing only **iron** medicinal product	product containing only **levorphanol** medicinal product
medical and surgical gastrointestinal system insertion **ileum** via natural or artificial opening endoscopic **infusion** device	medical and surgical gastrointestinal system revision **stomach** via natural or artificial opening endoscopic **other** device

Examples of False Positives Identified and False Negatives Not Identified

False Positives (Non-synonyms) Identified	
finding of **wrist** joint	finding of **knee** joint
malignant neoplasm of upper limb	malignant neoplasm of **muscle of** upper limb
skin **wound** of axillary fold	skin **cyst** of axillary fold
False Negatives (Synonyms) Not Identified	
hla antigen	human leukocyte antigen
pyelotomy	incision of renal pelvis treatment
routine cervical smear	screening for malignant neoplasm of cervix

Conclusion & Future Work

- Deep learning approach provides good performance in identifying synonymy among atoms.
- Adding source synonymy yields better precision and overall F-1 score.
- Adding hierarchical context trades precision for higher recall.
- Adding source synonymy, hierarchical context, and semantic group give an overall boost to accuracy and recall.

- **Limitation:** This approach does not address the *inter-concept* and *semantic type categorizations* *(other components in the UMLS Metathesaurus)*.

- **Future work:** How can the models be used in conjunction to complement the current lexical processing and human editors.

Acknowledgement

Dr. Olivier Bodenreider

Dr. Paul Fontelo

Dr. Vinh Nguyen

Rashmie Abeysinghe

Karan Luthria

NIH National Institutes of Health

NIH U.S. National Library of Medicine

References

- *Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research, 32(suppl_1), D267-D270.*

- *Mueller, J., & Thyagarajan, A. (2016, March). Siamese recurrent architectures for learning sentence similarity. In Thirtieth AAAI Conference on Artificial Intelligence.*

- *Wang, J., Yu, L. C., Lai, K. R., & Zhang, X. (2016, August). Dimensional sentiment analysis using a regional CNN-LSTM model. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 225-230).*

- *Nicosia, M., & Moschitti, A. (2017, November). Accurate sentence matching with hybrid siamese networks. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (pp. 2235-2238). ACM.*

- *Pontes, E. L., Huet, S., Linhares, A. C., & Torres-Moreno, J. M. (2018). Predicting the Semantic Textual Similarity with Siamese CNN and LSTM. arXiv preprint arXiv:1810.10641.*

Thank you!